Genet. Res., Camb. (2004), **84**, *pp.* 135–143. *With* 4 *figures.* © 2004 Cambridge University Press
DOI: 10.1017/S0016672304007190 Printed in the United Kingdom

Individual organisms as units of analysis: Bayesian-clustering alternatives in population genetics

JUDITH E. MANK* AND JOHN C. AVISE
Department of Genetics, University of Georgia, Athens, GA 30602, USA

(*Received 20 January 2004 and in revised form 20 August 2004*)

Summary

Population genetic analyses traditionally focus on the frequencies of alleles or genotypes in 'populations' that are delimited *a priori*. However, there are potential drawbacks of amalgamating genetic data into such composite attributes of assemblages of specimens: genetic information on individual specimens is lost or submerged as an inherent part of the analysis. A potential also exists for circular reasoning when a population's initial identification and subsequent genetic characterization are coupled. In principle, these problems are circumvented by some newer methods of population identification and individual assignment based on statistical clustering of specimen genotypes. Here we evaluate a recent method in this genre – Bayesian clustering – using four genotypic data sets involving different types of molecular markers in non-model organisms from nature. As expected, measures of population genetic structure (F_{ST} and Φ_{ST}) tended to be significantly greater in Bayesian *a posteriori* data treatments than in analyses where populations were delimited *a priori*. In the four biological contexts examined, which involved both geographic population structures and hybrid zones, Bayesian clustering was able to recover differentiated populations, and Bayesian assignments were able to identify likely population sources of specific individuals.

1. Introduction

Both in theory and in practice, the field of population genetics has traditionally focused on composite genetic properties of assemblages of individuals (Hartl & Clark, 1997). For questions relating to geographic population structure and gene flow as well as conservation genetics, populations are typically specified *a priori* and their allele or genotype frequencies are then analysed to reveal patterns of kinship, spatial structure, hybridization, etc. Examples of such approaches include the use of Wright's *F*-statistics (1921) or their derivatives, and population genetic distances exemplified by Nei's *D* (1987). Such composite parameters are then used to infer ecological or evolutionary forces at work in particular populations.

This conventional approach is not without shortcomings. One potentially serious difficulty involves the *a priori* assignment of individuals to specific populations, usually based on collection locales or phenotypes. This can introduce biases as well as dangers of circular reasoning, as individuals collected from specific locales may have originated from elsewhere. Thus traditional population genetic approaches can miss admixed or hybrid individuals, and related biological processes. Furthermore, information is lost in the process of compiling genotypes into allele frequencies across individuals. In allozyme or microsatellite surveys, for example, hundreds or even thousands of genotypes are averaged to generate estimates of population allele frequencies. Thus, immigrants might be lumped with natives, relatives lumped with non-relatives, etc., diminishing the power of any subsequent F_{ST} or distance-based analyses to dissect biological processes. Finally, in many traditional analyses, empirical population-genetic data are often interpreted in the context of unrealistic models (e.g. island models that assume that populations of equal size all exchange migrants at equal

* Corresponding author. e-mail: jemank@uga.edu

rates) and presumptions of genetic equilibria (e.g. between genetic drift and gene flow) that are violated routinely by real populations (Bossart & Prowell, 1998).

Thus, it would seem desirable to have an objective method for recognizing each individual's true population identity before initiating population genetic analyses. Several approaches have moved in this direction. Indeed, the view of individuals as operation units (OUs) was one of several key features in the phylogeographic revolution in population genetics (review in Avise, 2000). In particular, non-recombining haplotypes in the mitochondrial (mt) DNA genome permitted the provisional genealogical assignment of each individual to a particular matriline regardless of population membership as defined by other criteria such as collection locale or taxonomic assignment.

In principle, individuals might also be treated as OUs with regard to their nuclear genotypes. In an early example of this approach, Bowcock *et al.* (1994) summarized genotypic data at 30 microsatellite loci into a neighbor-joining tree whose external nodes were 148 individual humans from around the world. Despite only minor variations in allele frequency between regional populations, branches connecting individuals in the tree (based on percentages of alleles shared across loci) proved to reflect these people's geographic origins with considerable accuracy. Other examples in which individuals were treated as OUs, based on genotypic data from multiple nuclear loci, include molecular analyses of *Apis* honey bees (Estoup *et al.*, 1995), *Heterocephalus* mole-rats (O'Riain *et al.*, 1996) and *Odocoileus* deer (Blanchong *et al.*, 2002).

Some of the problems of traditional population genetics mentioned above (assigning individuals to populations, discriminating numbers of populations within a sample and drawing correct population boundaries) have also been addressed recently through Bayesian and related clustering methods as applied to genotypic data from individuals (Cornuet *et al.*, 1999; Dawson & Belkhir, 2001; Manel *et al.*, 2002; Pritchard *et al.*, 2000). For example, Bayesian clustering was used to study genetic structure in humans, and it was found that approximately 150 microsatellite loci were needed to attain sufficient relief in a likelihood topology to properly cluster 1056 individuals into their populations of origin (Rosenberg *et al.*, 2002). Bayesian methods have also been attempted in 'non-model' organisms (e.g. Cegaleski *et al.*, 2003, Mank *et al.*, 2004, Miller *et al.*, 2003), but these studies usually suffer from smaller sample sizes, availability of fewer information-rich markers, and generally less robust data sets. The net result of Bayesian analysis of this sort is lower signal-to-noise ratios in the genetic data, and shallower likelihood topologies for Bayesian clustering.

Here we further assess the utility of Bayesian population clustering by applying a popular version of this method (the program STRUCTURE; Pritchard *et al.*, 2000) to several natural populations for which we had genotypic data on individuals available. We purposefully chose examples (involving fire ants, a freshwater turtle, a catadromous marine fish and a hybrid treefrog population) to include organisms of diverse types, molecular markers with alternative modes of inheritance and biological questions of diverse etiology. Furthermore, these examples represent the kinds of molecular data often available for non-model species (i.e. those not extensively characterized at large numbers of genetic loci). Specifically, we were interested to learn whether Bayesian clustering could find relevant application in analysing genotypic data sets that are considerably smaller than those typically available for model organisms. For each example, we also compared results from Bayesian clustering with those gleaned from traditional frequentist approaches of population genetics.

2. Materials and methods

(i) *Data sets analysed*

We analysed four genetic data sets, detailed in Table 1, involving species with different dispersal syndromes and for which different types of molecular markers were available. The first data set involved 531 imported red fire ants (*Solenopsis invicta*) from nine collection sites in the introduced (US) range. Collections included multiple samples from both monogyne (single-queen) and polygyne (multi-queen) colonies in Georgia, and samples from a single monogyne and polygyne colony in Louisiana. Social form was determined independent of genetic data as per Ross & Shoemaker (1997). Individuals were genotyped with several co-dominant RAPD markers, allozymes and microsatellites (Ross *et al.*, 1999). These three classes of molecular markers were analysed both separately and combined in the current study.

The second data set consisted of haplotype data from mitochondrial restriction-site (RFLP) analyses of 67 southeastern mud turtles (*Kinosternon subrubrum* and *K. baurii*). Such mtDNA data presumably entail linkage disequilibrium inherent in the mitochondrial genome's non-recombinational mode of inheritance. We wanted to determine whether Bayesian analysis would reveal the same matrilineal clusters as did the conventional phylogeographic analysis that also treated those same individuals as OUs (Walker *et al.*, 1998).

The third data set, involving North Atlantic eels, was comprised of microsatellite genotypes at six loci from a total of 330 individuals from Europe and Iceland (*Anguilla anguilla*) and from North America

Table 1. *Characterization of data sets subjected herein to Bayesian analyses*

Data set and population	No. of samples	No. of allozyme markers	No. of micro-satellite markers	No. of co-dominant RAPD markers	No. of mtDNA RFLP markers	Nature of data
Solenopsis invicta Total	531	8	7	5	0	Genotypic
Georgia monogyne	133					
Georgia polygyne	149					
Louisiana	249					
Kinosternon Total	67	0	0	0	81	Haplotypic
Clade 1[a]	39					
Clade 2[a]	23					
Clade 3[a]	5					
Anguilla Total	330	0	6	0	0	Genotypic
A. anguilla (Iceland and Europe)	282					
A. rostrata (North America)	48					
Hyla Total	305	5	0	0	0	Genotypic
H. cinerea	104					
H. gratiosa	60					
Putative hybrids	141					

[a] As determined by original parsimony analysis in Walker *et al.* (1998).

(*A. rostrata*). Because North Atlantic eels are catadromous (they are born and spawn in the Sargasso Sea, but disperse to continental waters while juveniles), there is little genetic differentiation among rivers *within* either species. Phylogeographic mitochondrial analyses (Avise *et al.*, 1986) and to some extent frequentist microsatellite appraisals (Mank & Avise, 2003) yielded patterns of sharp genetic divergence between the two continental species but limited variation within either of them (see also Wirth & Bernatchez, 2001, 2003). In earlier analyses, there was also some possible evidence for interspecific hybridization, with individuals of hybrid ancestry perhaps travelling to Iceland along with other genetically pure *A. anguilla* specimens (Avise *et al.*, 1990).

The fourth data set was composed of 305 treefrogs typed at five allozyme loci (Lamb & Avise, 1986). The samples included two species (*Hyla cinerea* and *H. gratiosa*) and their hybrids collected from ponds near Auburn, Alabama. Previous work showed that fertile F_1 hybrids occur at this location and sometimes mate to produce backcross or later-generation hybrids that are not easily assigned by morphology alone to specific hybrid categories (Lamb & Avise, 1987). For current purposes, genetic data from this population were first Bayesian-analysed without prior assumptions about species membership, and then again in an additional analysis that distinguished pure *H. cinerea* from pure *H. gratiosa* but left all other individuals unspecified at the outset.

(ii) *Bayesian analyses*

For each data set, we used STRUCTURE 2.0 (Pritchard *et al.*, 2000) with a burn-in of 10 000 generations with 500 000 MCMC (Markov Chain Monte Carlo) generations, except in the case of the combined fire ant data set where we employed 100 000 burn-in generations with 1 000 000 MCMC generations. After initial analyses indicated that the 'admixture' model was a better fit to all our data sets, for each analysis we defined the maximum degree of admixture (α) among populations as $\alpha = 50$, and the proposal for updating as $\alpha = 2$. We chose these large values of α for several reasons. First, during preliminary test runs, higher α values helped the program reach convergence. Second, some of the data sets (on eels, treefrogs, and perhaps fire ants) indicated recent genetic admixture for the populations examined. Finally, higher α values in general put less constraint on the program during likelihood searches.

A typical Bayesian search resembles a maximum likelihood clustering approach when information on collection locales is ignored (Congdon, 2001) and the search of the topology is restricted to reasonable parameter values (uninformed or uniform priors). In this study, we began each data analysis with uniform priors. The Bayesian method can also incorporate information from data outside the experiment, such as collection locale or distinct phenotype that can shift the posterior distribution (Edwards, 1992). We also used informed priors in the current study, for comparison. However, when an informed prior is

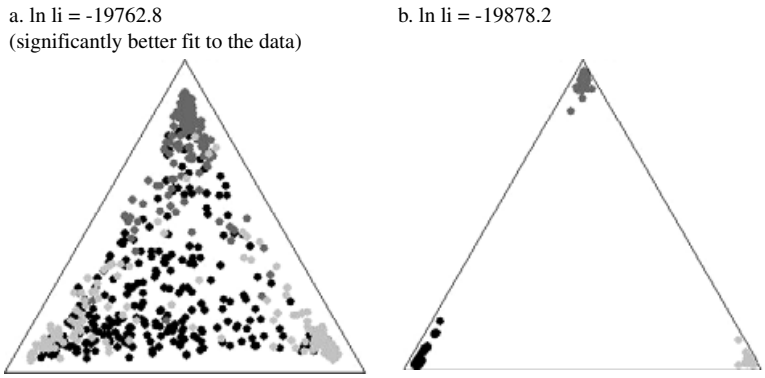

Fig. 1. Cluster plots summarizing results of Bayesian analyses of genetic data on fire ants. Each point is an individual. (*a*) Analyses based on uniform priors (no assumptions about populations or sites to which individuals belong); (*b*) analyses based on informed priors from population locale information. Black dots are individuals from Georgia polygyne nests, light grey dots are individuals from Georgia monogyne nests, and dark grey dots are specimens from Louisiana.

employed in Bayesian clustering, similar caveats arise as under conventional frequentist approaches in population genetics (i.e. regarding *a priori* assignment of individuals to populations). Nonetheless, analyses involving informed priors can be useful in searching for admixture zones or hybrid individuals after population clustering has been performed (Pritchard *et al*., 2000). Results from searches using both uniform and informed priors are illustrated as cluster plots that represent the genetic similarity of each individual to each inferred cluster. The most likely number of clusters (k) was computed by maximizing the posterior probability of k ($k=1$, $k=2$, …, $k=5$). Search outcomes were compared by likelihood ratio tests with one degree of freedom. Each analysis was repeated at least three times to test for convergence for each k.

The complexity of a likelihood topology is proportional to the number of populations and parameters inferred. Complex topologies are difficult for MCMC to explore effectively, so, based on preliminary analysis, it is sometimes necessary to substructure the data to reduce the number of clusters. For each data set examined here, Bayesian-identified populations were further analysed for substructuring, and triangular cluster plots were used to help visualize the three main populations that were identified. To determine the level of agreement between *a priori* and *a posteriori* assignments, Bayesian population assignment tests for individuals were then performed. Each test yields an *a posteriori* probability that a given specimen originated from each identified population, given that the population was deduced correctly. Values of F_{ST} for nuclear genotypic data, or Φ_{ST} for mtDNA haplotype data, were calculated with GenAlEx (Peakall & Smouse, 2001) both for the *a priori*-defined groups and for groups defined by individual assignment tests.

3. Results

(i) Solenopsis *ants*

Data from the three different classes of molecular markers (RAPDs, allozymes and microsatellites) from the nine collection sites were first run separately in STRUCTURE, with results in each case yielding evidence for three primary genetic clusters: monogyne ants in Georgia, polygyne ants in Georgia, and Louisiana ants. Because these different markers gave generally concordant patterns, they were then combined into one data set that was analysed using uniform priors, with results shown in Fig. 1*a*. Note the concentrations of dots (individuals) at the three tips of this cluster plot, but also the appearance of intermediate dots throughout the diagram. Thus, although genetic clusters emerged from the uniform priors analysis, they were far from comprehensive in including all individuals (perhaps due to inter-populational gene flow or incomplete lineage sorting from the ancestral conditions, for example). It is interesting that Bayesian clustering was generally able to distinguish the Georgia social forms, but not the Louisiana forms. This parallels the findings in Ross *et al*. (1999) and may be a result of limited sampling from the Louisiana populations.

Data were also analysed using the three groups listed above as informed priors, with results shown in Fig. 1*b*. Note that in this cluster plot, based on inferred genetic groups from the preliminary Bayesian assessment, all 531 individual ants (100%) now cluster tightly into the tips of the triangle (representing the Georgia monogyne, Georgia polygyne and Louisiana assemblages). Clearly, the use of informed priors greatly 'improved' the clustering (at the expense, however, of suffering from the potential flaw of assigning specimens to populations *a priori*).

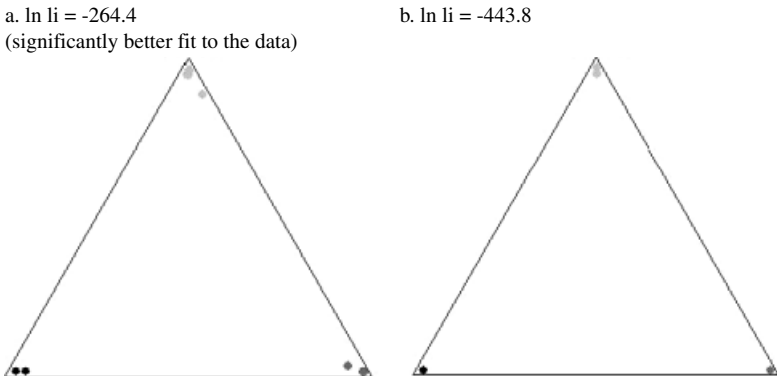

Fig. 2. Cluster plots summarizing results of Bayesian analyses of genetic data on mud turtles. Each point is an individual (although many individuals had the same mtDNA genotypes, so many dots are piled on top of one another). (*a*) Analyses based on uniform priors (no assumptions about populations or sites to which individuals belong); (*b*) analyses based on informed priors from population locale information. Black, light grey and dark grey dots are individuals from different clades as previously recognized in parsimony-based phylogenetic analyses (Walker *et al*., 1998).

The red imported fire ant in North America is a recently introduced species not likely to be in population-genetic or demographic equilibrium across its new range (Ross, 2001). This example illustrates the utility of Bayesian analyses in a data set with multiple collection locales and multiple classes of genetic markers. The inferred Bayesian clusters agree well with (and pictorially highlight) other previously published genetic analyses that have identified significant impacts for social behaviour (monogyny versus polygyny) as well as geography (Georgia versus Louisiana) on population genetic structure in this species (Ross & Shoemaker, 1997; Ross *et al*., 1999).

(ii) Kinosternon *turtles*

Bayesian clustering as applied to mtDNA haplotype data for *Kinosternon* freshwater turtles identified several population clusters that were entirely consistent with those identified in the original parsimony analysis of these data (Walker *et al*., 1998). To simplify and better visualize outcomes from the current analysis in cluster plots, we arbitrarily removed one clade (consisting of 16 turtles displaying mtDNA haplotype *subr*19) from the presentations. Fig. 2*a* and *b* show results of these Bayesian analyses based on uniform (collection locales unspecified) and informed (collection locales specified) MCMC runs. In both cases, all 67 turtles analysed in the data set (representing 25 distinct mtDNA haplotypes) grouped tightly into one or another of the three tips (distinct populations) in the cluster plot.

Furthermore, these groups were not invariably isomorphic with the collection locales. In other words, even in the Bayesian analyses that were informed by collection locality data, the maximum likelihood solution correctly grouped individuals by matrilineal rather than geographic proximity. This is apparently because the matrilineal signal from the haplotypes was strong enough to override any shift in the posterior probability distribution caused by the prior information. This example not only illustrates the utility of Bayesian approaches as applied to haplotype information from non-recombining genomes but also amplifies the previously recognized utility of mtDNA analyses that begin with individual organisms as OUs.

(iii) Anguilla *eels*

The Bayesian analyses with informed priors (collection locales) clearly identified three statistically significant genetic groupings of North Atlantic eels (Fig. 3*b*), coinciding with samples collected from North America, Europe and Iceland. Among these three assemblages, Icelandic eels were least tightly grouped genetically, with several specimens straying somewhat from the Icelandic position in the cluster plot. This pattern could be deemed consistent with previous evidence that some Icelandic specimens might be of hybrid ancestry (Avise *et al*., 1990), although other possibilities are by no means excluded.

However, in the uniform Bayesian analyses (those not informed by collection-locale priors), results from the microsatellite data were much less clear. Several of the eels from Iceland clustered at one apex, but most others did not, nor were European eels reliably demarcated from American eels in the cluster plots (Fig. 3*a*). We suspect this outcome to reflect forensic inadequacies of these particular microsatellite loci. This in turn might be due in part to homoplasy attributable to rapid interconversions between a limited number of allelic states (Goldstein & Schlötterer, 1999; Mank & Avise, 2003), or it might reflect a shallowness in the likelihood surface (thereby indicating the need for more microsatellite loci in

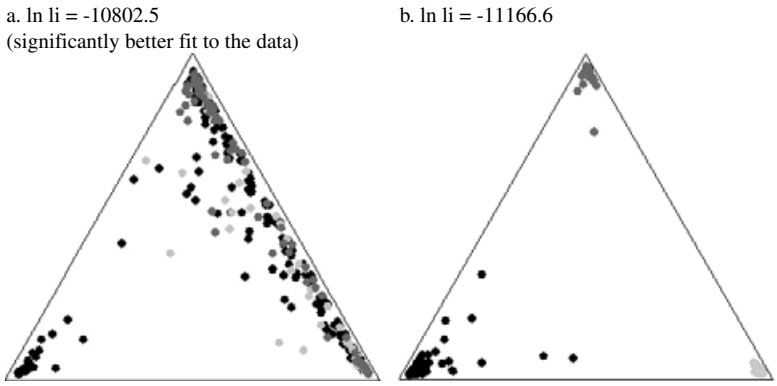

Fig. 3. Cluster plots summarizing results of Bayesian analyses of genetic data on eels. Each point is an individual. (*a*) Analyses based on uniform priors (no assumptions about populations or sites to which individuals belong); (*b*) analyses based on informed priors from locale information. Black, light grey and dark grey dots are individuals from Iceland, North America and Europe, respectively.

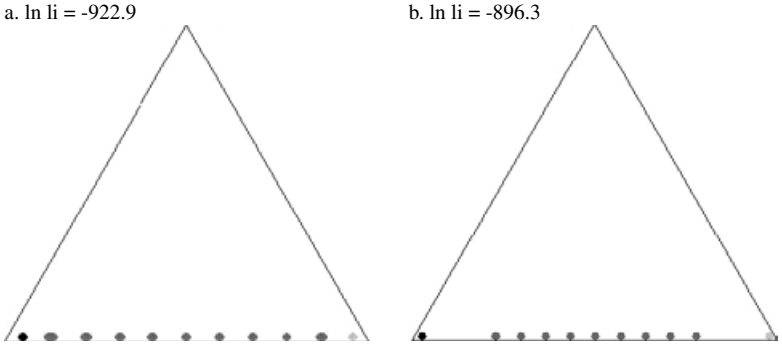

Fig. 4. Cluster plots summarizing Bayesian analysis of genetic data on *Hyla* treefrogs. (*a*) Analyses based on uniform priors (no assumptions about species or hybrids); (*b*) analyses based on informed priors (genetically 'pure' individuals assigned to two parental species). There are 11 genetic classes of individuals: one each for members of the two parental species and nine gradations reflecting combinations of parental alleles in hybrid classes. Individuals of *H. cinerea* are shown in black, *H. gratiosa* individuals in light grey and putative hybrids in dark grey (there are 305 individuals in these plots, but specimens in the same genetic class are stacked on top of one another).

the Bayesian analysis). We doubt that these results reflect biological realities about these populations, as analyses with other markers (mtDNA haplotypes as well as allozymes) showed quite clear genetic delineation between American and European eels (review in Avise, 2003), indicating that these two taxonomic species probably do represent two largely independent gene pools.

This example involving eels illustrates the danger of basing Bayesian (or any other) analyses on data from only a few nuclear loci, especially if these are compromised by homoplasy. It also illustrates the danger of placing too much confidence in the biological significance of genetic differences (even if statistically significant) between populations that were defined *a priori* by other criteria such as collection site. This problem is analogous to reading too much into small but statistically significant differences sometimes documented between *a priori*-defined demes in traditional population genetic analyses.

(iv) Hyla *treefrogs*

In the original genetic analyses by Lamb & Avise (1986), genotypes for pure (non-hybrid) individuals of the two treefrog species were homozygous for alternative, species-diagnostic alleles at each of the five allozyme loci surveyed. Each animal of supposed hybrid ancestry was either heterozygous at one or more loci, or else homozygous for different species-specific alleles at two or more loci, thus creating a total of nine distinct genetic classes of intermediates depending on how many alleles an individual carried from each of the two respective parental species.

Bayesian analysis was likewise able to discern parental and hybrid individuals with both uniform (Fig. 4*a*) and informed (Fig. 4*b*) priors: non-hybrids were positioned at two corners of the cluster plot, and all hybrids were intermediates between the parentals. These hybrid classes were few in number in this case, but a more nearly continuous distribution of

Table 2. *Comparisons of* a priori *and* a posteriori *outcomes in each of the four data sets. In two columns toward the left are shown percentages of individuals assigned by Bayesian likelihoods to populations defined* a priori *by geography or other external evidence, versus* a posteriori *Bayesian clusters defined by genotype. In the rightmost two columns are shown F_{ST} values for populations likewise defined* a priori *versus* a posteriori

Data set and population	*A posteriori* assignment[a] agrees with *a priori* assumption (%)	*A priori* assumption agrees with *a posteriori* assignment[a] (%)	*A priori* F_{ST}(±SD for genotypes, Φ_{ST} for haplotypes	*A posteriori* F_{ST} (±SD) for genotypes, Φ_{ST} for haplotypes
Solenopsis invicta Total	67·6	67·6	0·027 (0·0024)	0·040 (0·0020)
Georgia monogyne	93·2	70·7		
Georgia polygyne	55·0	55·7		
Louisiana	61·4	73·2		
Kinosternon Total	100	100	0·872	0·872
Clade 1[b]	100	100		
Clade 2[b]	100	100		
Clade 3[b]	100	100		
Anguilla Total	65·2	65·2	0·013 (0·00003)	0·025 (0·0050)
A. anguilla (Iceland and Europe)	62·1	93·9		
A. rostrata (North America)	83·3	27·2		
Hyla Total	100	100	1 (0·000)	1 (0·000)
H. cinerea	100	100		
H. gratiosa	100	100		
Putative hybrids	100	100		

[a] Based on most likely source population from inferred clusters.
[b] As determined by original parsimony analysis in Walker *et al.* (1998).

dots might have been present in the cluster plots if additional loci had been assayed.

4. Discussion

(i) *Individual assignment and measures of population differentiation*

Bayesian assignments of individuals to populations are based on the concept that specimens from the same true population will have genotypes more similar than individuals from different populations (Cornuet *et al.*, 1999), where genetic similarity acts as a proxy for genetic ancestry. For example, Bayesian methods should in principle assign immigrants and other non-natives to their respective populations of origin (if sampled), rather than to the recipient populations where they may have been collected. Assignment accuracy is of obvious importance to studies of genetic structure, phylogeography, and many questions in conservation genetics.

Bayesian assignment of individuals in the *Kinosternon* and *Hyla* data sets agreed perfectly with *a priori* assumptions about populations of origin based on collection locale. These *a priori* assumptions had been based on strong morphological or parsimony information. The population genetic separations were also deep (as shown by the high F_{ST} and Φ_{ST} values; Table 2). Thus, these two data sets were entirely transparent to both frequentist and Bayesian analyses, and this is reflected in the solid agreement between these two data treatments. The transparency of the *Kinosternon* data set is especially noteworthy due to the linked nature of the mtDNA markers utilized, meaning that the genetic appraisal in effect was based on a single 'gene'.

In the *Solenopsis* and *Anguilla* data sets, individual Bayesian assignments agreed less well with *a priori* assumptions about populations of origin (i.e. they 'correctly' assigned only 67·6% and 65·2% of the specimens, respectively). This relates to the fact that in either data set, geographic populations were only weakly differentiated genetically, perhaps due to histories of migration and gene flow, lack of complete lineage sorting from a polymorphic ancestral population, or occasional homoplasy (evolutionary convergence or reversals of state in the highly variable allelic markers examined; see Mank & Avise, 2003). However, it is not proper to conclude that Bayesian assignments are automatically correct and the *a priori* assignments are wrong. Any such conclusion would require absolute knowledge of the true genealogical histories of wild-caught individuals, a type of understanding that is rare if not unattainable for most natural populations. Thus, although Bayesian clustering may be a helpful tool in these situations, by itself it is of course not the ultimate arbiter of truth.

In general, Bayesian-derived estimates of population structure should be equal to or greater in magnitude than comparable estimates based on *a priori* assignment of individuals to geographic populations (because Bayesian assignments are based on genotypic clustering rather than external evidence such as geographic collecting locale). This expectation is supported in our data sets (Table 2). For *Kinosternon* and *Hyla*, the Φ_{ST} and F_{ST} values were identical in the *a priori* and *a posteriori* (likelihood) treatments, and such values were significantly higher in the Bayesian treatments of the *Solenopsis* and *Anguilla* data sets.

(ii) *Shortcomings of Bayesian clustering*

Bayesian techniques rely on MCMC searches to explore the likelihood space, but there are ambiguities in the literature as to how to direct the search. Number of parameters to estimate, shallowness of the likelihood topology (e.g. when relatively few data points are available or populations are very similar genetically), complexity of the parameter space and other characteristics can increase both the time to plateau (which should influence the desired burn-in length) and convergence on that plateau (which should influence chain length). These features vary by data set, and there is no simple diagnostic to choose these parameters nor to determine when chains have converged on the optimal solution (Cowles & Carlin, 1996). Depending on the algorithm used and the complexity and slope of parameter space, some MCMCs may have difficulty leaving the neighbourhood of an attractive solution that is not the global optimum (Chen *et al*., 2000). To protect against this, multiple MCMC searches should be performed.

Most studies of wild species are based on only a handful of markers. Although this might be interpreted as sufficient for meaningful signal in various frequentist population assessments, it can result in shallow likelihood topologies in Bayesian searches. Without using putative population assignments as priors, even extremely long MCMC searches do not always reach convergence. Using priors speeds the process, but may seriously bias the parameter space searches (as illustrated by our eel example) and thus provide little or no improvement over frequentist approaches. The use of priors also creates dilemmas as to the strength of prior evidence, and how heavily it should be weighted (Edwards, 1992). In cases of shallow likelihood topology where the signal from a given data set is insufficient to override *a priori* assumptions, the use of informed priors can result in assignments that merely recover the information given in the priors. In such instances, Bayesian approaches do little to remedy any underlying problems associated with *a priori* population recognition.

Another difficulty is that the most popular implementation of Bayesian clustering (STRUCTURE, the program employed here) has no explicit test for significant clustering. Although there are methods to determine the probability of the inferred number of clusters (k), and it is possible to compare results to those derived other likelihood-based methods (Dawson & Belkhir, 2001), there is currently no direct way (without resorting to frequentist perspectives such as F_{ST} analyses or an *ad hoc* likelihood test of the admixture model against a model of no admixture) to test whether observed clusters differ significantly from complete admixture, or whether different inferred clusters differ significantly from one another. A definitive significance test of inferred Bayesian clusters against a null model of complete admixture (for example) would be helpful.

(iii) *General implications*

Our analyses and others like them indicate that Bayesian clustering methods can be an informative adjunct to traditional population genetic approaches. They can circumvent requirements to define populations *a priori* and to average genotypic data into population allele frequencies, and they can ameliorate the problem of testing whether collection locales are fair predictors of population membership. However, when based on only a relatively few nuclear marker loci (especially if there is suspected homoplasy, as exemplified by the eel data set), Bayesian clustering may have limited applicability for several reasons, including the possibility of a shallow topology in the likelihood surface. Haplotype information from mtDNA (as in the phylogeographic data set on turtles), or from multi-locus individual genotypes (as illustrated by the hybrid-zone data set on treefrogs) appear to be amenable to Bayesian analysis, but in those cases more traditional analyses already treated individuals as singletons or OUs, thus obviating much of the special rationale otherwise reserved for Bayesian methods. Combined data sets that employed many marker types (as illustrated by the data sets on fire ants) would appear to offer some of the most promising opportunities for Bayesian clustering approaches in providing genuinely new and helpful approaches for population identification and individual assignments.

This work was supported in part by a University-Wide Fellowship from the University of Georgia and National Institutes of Health training grant to J. E. M., and by funds from the Pew Foundation to J. C. A. We thank K. G. Ross and D. Walker for the use of their data sets and useful comments on the manuscript. R. Kuzoff, E. Dakin, M. Mackiewicz, N. Leahy and C.-H. Kuo also offered helpful comments to improve the manuscript. D. P. Brown provided much appreciated computer assistance.

References

Avise, J. C. (2000). *Phylogeography: The History and Formation of Species*. Cambridge, MA: Harvard University Press.

Avise, J. C. (2003). Catadromous eels of the North Atlantic: a review of molecular genetic findings relevant to natural history, population structure, speciation, and phylogeny. In *Eel Biology* (ed. K. Aida). Berlin, Heidelberg, New York: Springer.

Avise, J. C., Helfman, G. S., Saunders, N. C. & Hales, L. S. (1986). Mitochondrial DNA differentiation in North Atlantic eels: population genetic consequences of an unusual life history pattern. *Proceedings of the National Academy of Sciences of the USA* **83**, 4350–4354.

Avise, J. C., Nelson, W. S., Arnold, J., Koehn, R. K., Williams, G. C. & Thorsteinsson, V. (1990). The evolutionary genetic status of Icelandic eels. *Evolution* **44**, 1254–1262.

Blanchong, J. A., Scribner, K. T. & Winterstein, S. R. (2002). Assignment of individuals to populations: Bayesian methods and multi-locus genotypes. *Journal of Wildlife Management* **66**, 321–329.

Bossart, J. L. & Prowell, D. P. (1998). Genetic estimates of population structure and gene flow: limitations, lessons, and new directions. *Trends in Ecology and Evolution* **13**, 202–206.

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R. & Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–457.

Cegaleski, C. C., Waits, L. P. & Anderson, N. J. (2003). Assessing population structure and gene flow in Montana wolverines (*Gulo gulo*) using assignment-based approaches. *Molecular Ecology* **12**, 2907–2918.

Chen, M.-H., Shao, Q.-M. & Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Analysis*. Berlin, Heidelberg, New York: Springer.

Congdon, P. (2001). *Bayesian Statistical Modeling*. New York: Wiley.

Cornuet, J.-M., Piry, S., Luikart, G., Stoup, A. & Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**, 1989–2000.

Cowles, M. K. & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* **91**, 883–904.

Dawson, K. J. & Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research* **78**, 59–77.

Edwards, A. W. F. (1992). *Likelihood*. Baltimore, MD: Johns Hopkins University Press.

Estoup, A., Garnery, L., Solignac, M. & Cornuet, J.-M. (1995). Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics* **140**, 679–695.

Goldstein, D. B. & Schlötterer, C. (eds.) (1999). *Microsatellites: Evolution and Applications*. Oxford: Oxford University Press.

Hartl, D. & Clark, A. G. (1997). *Principles of Population Genetics*. Sunderland, MA: Sinauer.

Lamb, T. & Avise, J. C. (1986). Directional introgression of mitochondrial DNA in a hybrid population of tree frogs: the influence of mating behavior. *Proceedings of the National Academy of Sciences of the USA* **83**, 2526–2530.

Lamb, T. & Avise, J. C. (1987). Morphological variability in genetically defined categories of anuran hybrids. *Evolution* **41**, 157–165.

Manel, S., Berthier, P. & Luikart, G. (2002). Detecting wildlife poaching: identifying the origin of individuals with Bayesian assignment tests and multilocus genotypes. *Conservation Biology* **16**, 650–659.

Mank, J. E. & Avise, J. C. (2003). Microsatellite variation and differentiation in North Atlantic eels. *Journal of Heredity* **94**, 310–314.

Mank, J. E., Carlson J. C. & Brittingham, M. C. (2004). A century of hybridization: decreasing genetic distance between American black ducks and mallards. *Conservation Genetics* **5**, 394–403.

Miller, C. R., Adams, J. R. & Waits, L. P. (2003). Pedigree-based assignment tests fir reversing Coyote (*Canis latrans*) introgression into the red wolf (*Canis rufus*) population. *Molecular Ecology* **12**, 3287–3301.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

O'Riain, M. J., Jarvis, J. U. M. & Faukes, C. G. (1996). A dispersive morph in the naked mole-rat. *Nature* **380**, 619–621.

Peakall, R. & Smouse, P. E. (2001). GenAlEx V5: genetic analysis in Excel. Population genetic software for teaching and research. Australian National University, Canberra, Australia. http://www.anu.edu.au/BoZo/GenAlEx/

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. & Feldman, M. W. (2002). Genetic structure of human populations. *Science* **298**, 2381–2385.

Ross, K. G. (2001). How to measure dispersal: the example of fire ants. In *Causes, Consequences, and Mechanisms of Dispersal at the Individual, Population, and Community Level* (ed. J. Clobert, E. Danchin, J. D. Nichols & A. Dhondt), chapter 1.3. Oxford: Oxford University Press.

Ross, K. G. & Shoemaker, D. D. (1997). Nuclear and mitochondrial genetic structure in two social forms of the fire ant *Solenopsis invicta*: insights into transitions to an alternate social organization. *Heredity* **78**, 590–602.

Ross, K. G., Shoemaker, D. D., Krieger, M. J. B., DeHeer, C. J. & Keller, L. (1999). Assessing genetic structure with multiple classes of molecular markers: a case study involving the introduced fire ant *Solenopsis invicta*. *Molecular Biology and Evolution* **16**, 525–543.

Walker, D., Moler, P. E., Buhlmann, K. A. & Avise, J. C. (1998). Phylogeographic pattern in *Kinosternon subrubrum* and *K. baurii* based on mitochondrial DNA restriction analyses. *Herpetologica* **54**, 174–184.

Wirth, T. & Bernatchez, L. (2001). Genetic evidence against panmixia in the European eels. *Nature* **409**, 1037–1040.

Wirth, T. & Bernatchez, L. (2003). Decline of North Atlantic eels: a fatal synergy? *Proceedings of the Royal Society of London, Series B* **270**, 681–688.

Wright, S. (1921). Systems of mating. *Genetics* **6**, 111–178.